Gemma 3 выделяется своей производительностью. Google утверждает, что она обходит такие известные модели, как DeepSeek V3 и OpenAI o3-mini, особенно в задачах, требующих анализа текста, изображений и коротких видео. Её «контекстное окно» — объём данных, который модель может обработать за раз, — достигает 128 тысяч токенов. Для примера, это эквивалентно книге на 200 страниц.

При этом модель настолько эффективна, что 27-миллиардная версия работает на одном графическом процессоре (GPU), таком как NVIDIA H100, а не на целой группе серверов.

Особенность Gemma 3 — её открытость. В отличие от закрытых моделей вроде ChatGPT, код Gemma доступен всем через платформы Google AI Studio, Hugging Face и Kaggle. Модель также поддерживает более 140 языков.